DOI:10.16016/j.1000-5404.2004.04.014
第26卷第4期 第三军医大学学报 Vol. 26, No. 4
2004年2月 ACTA ACADEMIAE MEDICINAE MILITARIS TERTIAE Feb. 2004


文章编号:1000-5404(2004)04-0317-03

论著

# 基因表达聚类结果的信息熵评价方法

易 东[1],杨梦苏[2],李辉智[3],黄明辉[2],王文昌[1] (1第三军医大学预防医学系卫生统计学教研室,重庆 400038;2香港城市大学基因组科技应用研究中心,香港 00852;3西南政法大学刑侦学院电子技术教研室,重庆 400031)

提 要:目的 探讨基因表达数据聚类结果信息熵评价方法,为合理地选择聚类方法提供一种定量评价依据。方法 采用 *entropy* 信息熵方法,考察常用的六种聚类方法所得到的分类结果与部分已知功能基因分类之间的符合程度,并将该指标作为一种评价依据。结果 将该方法应用于Lyer的血清刺激表达数据集聚类结果的评价,给出了六种聚类方法的 *entropy* 图。结论 本研究首次提出用熵理论评价聚类结果,并观察到同一数据集由于功能聚类信息的不同而引起的评价结果的差异。并且结果显示SOM和模糊聚类算法对分析处理Lyer的血清刺激的基因表达数据更为合理。

关键词:基因表达;聚类分析;评价; *entropy* 信息熵
中图法分类号:R195.1;R394-33 文献标识码:A

# A novel method for evaluation of clustering results for gene expression data based on *entropy* theory

YI Dong[1], YANG Meng-su[2], LI Hui-zhi[3], HUANG Ming-hui[2], WANG Wen-chang[1] ([1]Department of Health Statistics, College of Preventive Medicine, Third Military Medical University, Chongqing 400038, China; [2]Applied Research Center for Genomics Technology, City University of Hong Kong; [3]Department of Electronic Technology, College of Criminal Investigation, Southwest China University of Politics and Law, Chongqing 400031, China)

**Abstract**: Objective To establish a systematic framework for the selection of the best clustering algorithm and to provide an *entropy* evaluation method for clustering analysis of gene expression data. Methods Based on information theory, *entropy* is used to measure the consistency between the clustering results from six algorithms and the known and validated functional classifications. Results In this study, we applied the *entropy* method for Lyer's gene expression data. Six *entropy* curves of clustering algorithms were obtained. Conclusion According to the curve of *entropy*, both SOM and Fuzzy clustering methods show the highest ability to cluster on these two datasets.

**Key words**: gene expression; clustering method; evaluation; *entropy*

基因表达数据聚类分析的目的是通过表达变化趋势的相似性,去寻找基因的功能的相似性[1,2]。因此,聚类结果的评判应考虑是否和部分基因功能聚类相符合?目前,基因聚类结果评判方法主要基于数据结构,从而对聚类的质量进行评价[3]。但是,以上方法完全基于类里的数据结构对聚类的质量进行评价,其优点是可以不需要数据任何相关的背景知识,即可对类的质量做出评价。不过这也是它的缺点所在,因为从常识来看,基因芯片可以对成千上万个基因表达进行平行分析,我们不可能对所有的基因都一无所知,事实上往往是知道一部分基因的功能,并希望通过这些已知基因来认识未知基因,或通过基因的已知功能来认识未知基因的功能。对这些信息弃之不用,势必导致信息的丢失,影响评价的精确性。正是基于这一点,本研究探讨采用 *entropy* 信息熵,描述聚类分析结果与部分已知功能基因分类的符合程度,从而提供一种新的评价指标,以期提高聚类结果选择的准确性和合理性。

## 1 数据集的准备

Lyer的数据集[4]是尝试研究纤维原细胞对血清刺激表达的反应,利用cDNA微阵列观察大约8 600个基因,以获得隐藏在这些反应背后即时转录程序。对来自新生包皮的纤维原细胞进行原代培养,无血清处理,诱导其维持稳定状态48 h,然后用加有10%的新生胎牛血清刺激表达。用DNA微阵列杂交,检测8 613个基因在刺激后15 min至24 h,12个时间点mRNA水平相对于0 min(无FBS刺激)的变化。他们本来的目的是把这个试验当作一个模板,来研究细胞从$G_0$期到增生期的控制机制。但作为一个基因组水平的表达谱试验,对成千上万个不同基因的考察,为研究所有实际发生的过程打开了一扇窗户,而不仅仅是研究细胞从$G_0$期到增生期的控制机制。在文献[4]中,选择了一个含有517个基因的子集进行聚类分析。本研究将对6种常用的聚类方法在该子集数据上的聚类结果进行分析,采用的部分已知基因分类情况见表1、2。

作者简介:易 东(1964-),男,重庆市人,博士,教授,主要从事生物信息学方面的研究,发表论文30余篇。电话:(023)68752344

收稿日期:2002-07-31;修回日期:2003-02-20

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表1 517个基因集合中基于信号传导功能的部分基因分类

**Tab 1 Classification of a part of 517 genes based on signal transduction**

| Signal transduction | Immediate-early transcription factors | Other transcription |
|---|---|---|
| Protein tyrosine kinase Ror1 | ATF3 | DB1 |
| TGFBR3 | C-FOS | HIP116 |
| Ndr protein kinase | EGRP1 | Meis1 (mouse) homolog |
| Gem GTPase | PUTATIVE DNA BINDING PROTEIN | LDB1 |
| EST Highly similar to OPIOID BINDING PROTEIN | A20 | AH-receptor |
| EDG1 | HIV-1 PROTEIN 2 | HEAT SHOCK FACTOR PROTEIN 2 |
| NET1 | ETS-2 | DP2 |
| SERINE/THREONINE-PROTEIN KINASE SGK | ETS-2 | ERF-2 |
| SERINE/THREONINE-PROTEIN KINASE SGK | DEC1 | forkhead protein FREAC-2 |
| MAP KINASE PHOSPHATASE-1 | TGF-beta inducible early protein (TIEG) | ERF-2 |
| MAP KINASE PHOSPHATASE-1 | MINOR | |
| MAP KINASE PHOSPHATASE-1 | Jun B | |
| MAP KINASE PHOSPHATASE-1 | CPBP | |
| MAP KINASE PHOSPHATASE-1 | | |

表2 517个基因集合中基于伤口愈合功能的部分基因分类

**Tab 2 Classification of a part of 517 genes based on wound healing**

| Cell cycle and proliferation | | Coagulation and homeostasis | Tissue remodeling | Cytoskeletal reorganization |
|---|---|---|---|---|
| p57Kip | CCNA | THBD | CALLA/CD10 | SPTBN1 |
| p27Kip | CCNA | Factor Ⅲ | FBN2 | EST highly similar t-adducin |
| CDK6 inhibitor | CDC2 | Factor Ⅲ | FMOD | CALD1 |
| WEE1-like protein kinase | CKS2 | Endothelin 1 | LAMA2 | Desmophkin Ⅰ and Ⅱ |
| Pre-b-cell leukemia | CCNB1 | PLAUR | Alpha-1 type 3 collagen | Vimentin |
| hSIAH1 | CDK7 | PAI 1 | TIMP3 | EDG1 |
| PCNA | EST AA016305 | PAI 1 | Aminopeptidase N | NET1 |
| PCNA | ID3 | PAI2 | Elastin | |
| DNA topoisomerase | ID2 | TFPI2 | PLOD2 | |
| Ⅱ | CTPS | TFPI2 | CDH2 | |
| DNA topoisomerase | Cyclin D1 | | Furin | |
| Ⅱ | MK167 | | CTGF | |
| Madp2 | | | | |
| CENP-F | | | | |
| RRM1 | | | | |
| RRM2 | | | | |

## 2 *Entropy* 评价方法[5]

下面，我们将介绍基于信息熵 *entropy* 原理，描述聚类结果与功能分类的吻合度。

### 2.1 集合 $S$ 的 *entropy*

假如 $S$ 是一任意集合，$freq(C_i, S)$ 代表 $S$ 里属于集合 $C_i$ 元素的个数，用 $|S|$ 代表系列 $S$ 里的样本数。频率为 $\frac{freq(C_i, S)}{|S|}$，这样可得到 $C_i$ 所含有的信息量为：

$$-\log_2\left(\frac{freq(C_i, S)}{|S|}\right) \quad (1)$$

若将 $S$ 划分为：$C_1, C_2 \cdots, C_n$，那么可定义 $S$ 的 *entropy* 为：

$$info(S) = -\sum_{i=1}^{k}\frac{freq(C_i, S)}{|S|}\times \log_2\left(\frac{freq(C_i, S)}{|S|}\right) \quad (2)$$

### 2.2 分类实验 $X$ 和已知分类 $T$ 的符合度 $gain_ratio(X)$

已知分类用 $T$ 表示：$T_1, T_2, \cdots, T_n$。在实验分类 $X$ 将 $T$ 重新划分 $n$ 类后，其 *entropy* 的计算为

$$info_x(T) = \sum_{i=1}^{n}\frac{|T_i|}{|T|}\times info(T_i) \quad (3)$$

$$gain(X) = info(T) - info_x(T) \quad (4)$$

式(4)可以表达试验分类 $X$ 对 $T$ 分类的符合程度，其值越大，则符合程度越好。

同时，由于 $T$ 的分类数 $n$ 增大时，$gain(X)$ 会自然增大，所以排除该因素的影响，即除以 $split\ info(X)$ 消除分类数 $n$ 的影响。

$$split\ info(X) = -\sum_{i=1}^{n}\frac{|T_i|}{|T|}\times \log_2\left(\frac{|T_i|}{|T|}\right) \quad (5)$$

标准化公式为：

$$gain_ratio(X) = gain(X)/\ split\ info\ (X) \quad (6)$$

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

该公式能表达试验分类 $X$ 对实际 $T$ 分类的符合程度, 所获得的信息熵越高表示其符合程度就越好, 如 $gain_ratio(X)=1$ 时, 表示理论分类和实际分类完全吻合。因此, 该指标可以描述不同聚类结果与部分已知基因功能聚类的吻合程度。

### 2.3　分类方法

我们将用 *gain_ratio*(*X*)法对 6 种常用的聚类方法在 Lyer 的血清刺激表达数据集合上的聚类效能进行评价。这 6 种聚类方法分别为: K-means 方法、模糊聚类 Fuzzy、自组织映射 SOM、Hierarchical 谱聚类中的三种算法 average_link、single_link、complete_link。

## 3　结果

### 3.1　基于信号传导功能分类的评价结果

由上述基本原理可知, 对 *gain_ratio*(*X*)变化趋势图的判断标准应为: 当曲线位置越高, 则说明该聚类结果越合理。

图 1 显示的是 6 种算法的 *entropy* 变化曲线图。总体上来看, SOM 方法的聚类效果最好, Fuzzy 聚类法和 SOM 聚类次之, complete、single、average 聚类效果较差, K-means 方法介于之间。并且大多数算法的 *entropy* 曲线在聚类数为 11 附近有一个峰值, 该结果提示: SOM 方法和 Fuzzy 聚类方法适用于 Lyer 的血清刺激表达数据集, 并且最佳聚类数在 11 附近。

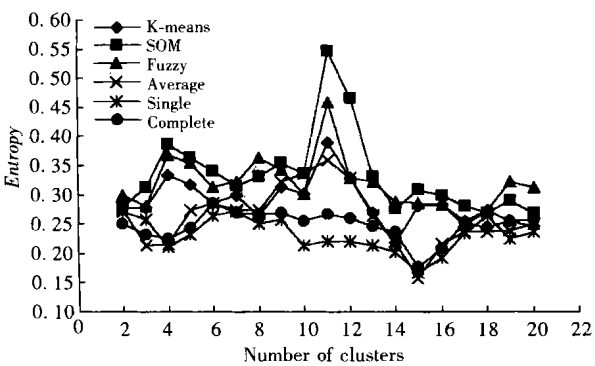



图 1　基于信号传导功能 ***Lyer*** 血清刺激表达数据的 ***gain_ratio***(***X***)变化趋势图

**Fig 1　*Entropy* resulted from the Lyer' s raw data and clustering data of fibroblasts response process**

### 3.2　基于伤口愈合功能分类评价结果

下面, 我们针对同一数据集, 采用另一外部标准, 即成纤维细胞基因在伤口愈合中的功能分类(如表 2 所示)来计算 *entropy*, 从而对 6 种聚类算法进行评判, 并对不同外部标准对算法的影响加以比较。

和图 1 不同的是图 2 的 *entropy* 是基于另外一个外部标准计算的。对 *entropy* 曲线进行比较可以发现 SOM 聚类效能最高, Fuzzy 聚类其次, 其它 4 种算法没有较大的差别。而且, *entropy* 曲线在聚类数为 13 附近呈现一个峰值, 这就提示最佳聚类数应在 13 附近。

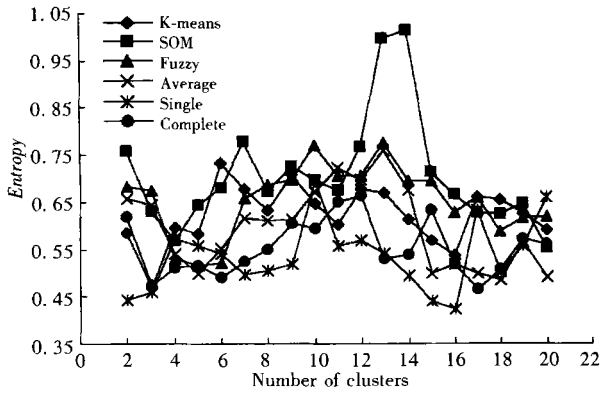



图 2　基于伤口愈合 **Lyer** 血清刺激表达数据的 ***entropy*** 变化趋势图

**Fig 2　*Entropy* resulted from the Lyer' s raw data and clustering data of fibroblast roles in the physiology of wound healing**

## 4　讨论

本研究提出利用部分已知基因功能分类, 采用信息熵方法对聚类结果进行评价。建立了一种新的聚类结果评价标准。该方法的思想来源于实际数据处理中出现的理论分类与部分功能分类的不吻合问题。因此, 本评价模型可以为实际工作者提供一种选择合理聚类结果的方法, 而且具有很强的可操作性。

从以上 6 种聚类算法在 2 个数据集的结果显示: ①SOM 聚类方法和 Fuzzy 聚类方法, 在此 3 个数据集上有相对高的聚类效能, K-means 方法具有中等聚类能力, average 聚类、single 聚类和 complete 聚类效果相差不大。因此, 我们推荐使用 SOM 聚类和 Fuzzy 聚类方法来分析基因表达数据; ②由图 1 和图 2 的 *entropy* 曲线图结果揭示了一个有趣的事实, 采用不同的外部标准, 即使对原始数据使用同一算法, 也会产生不同的最佳聚类结果。这也充分证实了加上外部评判标准的重要性, 也是我们推荐在聚类原始数据之前要考虑其功能分类的原因。

### 参考文献:

[1] Geschwind D H. Sharing gene expression data: an array of options[J]. Nat Rev Neurosci, 2001, 2(6): 435－438.

[2] 易　东, 杨梦苏, 李辉智, 等. 基于伪 F 统计量的模糊聚类方法在基因表达数据分析中的应用[J]. 中国卫生统计, 2002, 19(3): 146－151.

[3] Yeung K Y, Haynor D R, Ruzzo W L. Validating clustering for gene expression data[J]. Bioinformatics, 2001, 17(4): 309－318.

[4] Lyer R V. The transcription program in the response of human fibroblasts to serum[J]. Science, 1999, 283(1): 83－87.

[5] Mangasarian O L. Mathematical programming in data mining[J]. Data Mining and Knowledge Discovery, 1997, 20(1): 183－201.

(编辑　张大春)

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net